

Visualization for Knowledge Discovery in Database

S. O. Rezende, R. B. T. Oliveira, L.C. M. Félix & C. A. J. Rocha
Department of Computer Science and Statistics
Institute of Mathematical and Computer Sciences
University of São Paulo
Av. Dr. Carlos Botelho, 1465 – CP. 668 – CEP 13560-970
São Carlos – SP - Brazil
Email: {solange,robson,lmolina,crocha}@icmsc.sc.usp.br

Abstract

The ever larger interest of the companies in accompanying new processing technologies and storage of data, as well as using the information as a large patrimony, has motivated several researches for the study of the process of transformation of this data into knowledge, which provides an intelligent aid to decision-making. In this context, the process of Knowledge Discovery in Databases (KDD) arises as a technology which can help with the knowledge search in the data. This search can be accomplished with the aid of visualization techniques, which can facilitate the understanding, on the part of human *Analysts*, of the knowledge extracted from data. This can be accomplished by identifying structures, characteristics, tendencies, anomalies and relationships among the data. These techniques frequently offer mechanisms that facilitate the search of patterns/models from data bases.

1. Introduction

In the last decade, three important factors are leading to the growth of the capacity of generating and collecting data. First, the readiness of technologies that offer a greater storage power and data processing at a lower cost; second, the accumulation of data at a growing rate; and third, the introduction of new groups of methods developed by the computer science community for the data processing Decker[2], Li[9].

The ever larger interest of the public and private corporations in accompanying these new processing technologies and storage

techniques, as well as using the information as a great patrimony, has been addressed in much research into the process of transformation of this data into knowledge, which provides an intelligent aid in decision-making.

The transformation of data into knowledge has been using mostly manual methods for data analysis and interpretation, which makes the process of pattern extraction of databases too expensive, slow and highly subjective, as well as unthinkable if the volume of data is huge.

The interest in automating the analysis process of great volumes of data has been fomenting several research projects in an emergent field called Knowledge Discovery in Databases Fayyad[3]. KDD is the process of knowledge extraction from great masses of data with the goal of obtaining meaning and consequently understanding of the data, as well as to acquire new knowledge. This process is very complex because it consists of a technology composed of a group of mathematical and technical models of software that are used to find patterns and regularities in the data Decker[2].

The search for these patterns, as well as the understanding of the results from all the phases that compose the KDD process, can be accomplished with the aid of visualization techniques. These techniques facilitate the understanding, on the part of the human *Analysts*, of the knowledge extracted from the data, which can be accomplished by identifying structures, characteristics, tendencies, anomalies and relationships in the data.

This article shows some of the visualization techniques that can be applied to the stages of the KDD process in a specific collection of data. The paper is organized in the following way: Section 2 presents the characteristics of the data collection used in this work. Section 3 presents some basic concepts and related works involving the KDD process, the data visualization and the tools used in the study of cases. In Section 4 the visualization techniques used in each one of these phases are presented, considering a case study from the Program of Genetic Improvement of the Nelore Breed. Finally, in Section 5, the conclusions of this article are presented.

2. Data Collection

The Program of Genetic Improvement of the Nelore Breed (PMGRN – Programa de Melhoramento Genético da Raça Nelore) started in June 1988, with the union of the cattle breeders, looking for modern technologies and easy applications to cattle breeding, with the

researchers of the Genetics Department of the Medicine School (DGFM), of the University of São Paulo. Together, they defined goals for increasing the national herd productivity.

In October 1994, the use of the most advanced technology available on the international market for predicting the genetic value began, in the form of Expected Differences in Progeny (DEP – Diferença Esperada da Progenie) of the animals which were a part of the program: the Animal Model. DEP are predicted for weights and scrotum perimeters standardized for different ages.

The DEP is used to compare the genetic merit of animals for several characteristics and predict the genetic transmission ability of an animal evaluated as a progenitor. In PMGRN, the following DEP were considered:

- DEP for direct effect on the weight, which predicts the difference in expected kilograms, due to the genes for growth passed on by its father. DEPs were used for weights at 120, 240, 365 and 550 days of age;
- DEP for maternal effect, which predicts the weight difference of the products (children) of the daughters of the appraised animal, due to the differences in the maternal ability (production of milk) presented by them. DEPs was used for scrotum perimeters measured at 365 and 550 days of age;
- DEP for direct effect on the scrotum perimeter, which predicts the ability of the animal for transmitting genes of testicular growth to its progeny. DEPs was used for maternal effect at 120 days of age.

The value of an animal is not determined by an isolated characteristic, but by the combination of the characteristics of largest economic interest, for which it transmits genes to its progeny. Thus, the Total Genetic Merit (MGT – Mérito Genético Total) is the index that gathers several predictions into a single value trying to ponder them in a more adapted way for the maximization of the genetic progress and, consequently, the productivity.

3. Basic Concepts and Related Works

To extract knowledge from databases, it is essential that the *Expert* follows some steps or basic stages in order to find a path from the raw data to the desired knowledge. The KDD process organizes these stages in a sequential and iterative form. In this way, it would be interesting if the obtained results of these steps were analyzed in a more interactive and friendly way, seeking a better evaluation of these results. The

representation of these results in a graphic form gives a larger safety and comfort for analyzing and evaluating the results. These graphic representations or visualizations are obtained by using tools adapted for such a purpose. Thus, it is more natural to use tools that are associated to the KDD process.

In this section we comment some basic concepts and tools used in this work. The interested reader can find a more thorough description in Félix[5], Inmon[6], Kerber[7], Li[9], Oliveira[11].

3.1. Knowledge Discovery in Database

The process of knowledge extraction from databases combines methods and statistical tools, machine learning and databases to find a mathematical and/or logical description, which can be eventually complex, of patterns and regularities in data Fayyad[3], Mannila[10].

The knowledge extraction from a large amount of data should be seen as an interactive and iterative process, and not as a system of automatic analysis. In this way, we cannot simply expect an extraction of useful knowledge by submitting a group of data to a “black box” Mannila[10].

The interactivity of the KDD process refers to the greater understanding, on the part of the users of the process, of the application domain. This understanding involves the selection of a representative data subset, appropriate pattern classes and good approaches to evaluating the knowledge. For a better understanding the functions of the users that use the KDD process, in this paper users are divided in three classes: (a) *Domain Expert*, who should possess a large understanding of the application domain; (b) *Analyst*, who executes the KDD process and, therefore, he should have a lot of knowledge of the stages that compose this process and (c) *Final User*, who does not need to have much knowledge of the domain. Frequently, the *Final User* uses knowledge extracted from the KDD process to aid him in a decision-making process.

The success of the KDD process depends partly on the interaction among users. It is not probable that the *Analyst* will find useful knowledge from the data without the guarantee of the *Expert* as to what would be useful for a specific domain. Besides this, the interactivity of the process requires that the *Final User* and the *Expert* have an effective participation in the choices and decisions during the process Fayyad[3].

Knowledge discovery from data can be understood as a process that contains, at least, the steps of application domain understanding,

A representative outline containing all these steps is illustrated in Figure 1. The KDD process begins with the understanding of the application domain, considering aspects such as the objectives of the application and the data sources. Next, a representative sample (e.g. using statistical techniques) is removed from database, preprocessed and submitted to the methods and tools of the Data Miningⁱ stage with the objective of finding patterns/models (knowledge) in the data. This knowledge is then evaluated as to its quality and/or usefulness, so that it can be used to support a decision-making process. It should be emphasized that, in spite of the visualization tools being used mostly in the knowledge evaluation step, they have great relevance in understanding and evaluating the results of each stage, especially for the *Final User*.

It is important to notice that, because it is an iterative process, the KDD steps are not tight, that is, the correlation among the techniques and methods used in the several stages is considerable, to the point of the occurrence of a small change in one of them affecting substantially the success of the whole process. In this way, the results of a certain stage can change any of the previous stages or even make it necessary restart the whole process Fayyad[4].

3.2. Data Visualization

Data visualization products are being seen as an indispensable tool for KDD due to the increase in the analysis capacity and to the interpretation of the results. Data visualization can be applied as an initial and exploratory phase of Data Mining to identify the domains of interest. An Analyst can build a number of exhibitions of data to determine the most significant graph. The number and type of exhibitions of data that are being generated depend on the knowledge of the Analyst. Graphical visualizations are interesting to determine appropriate statistical data for mining, so that the data visualization tools can then be used to present the results.

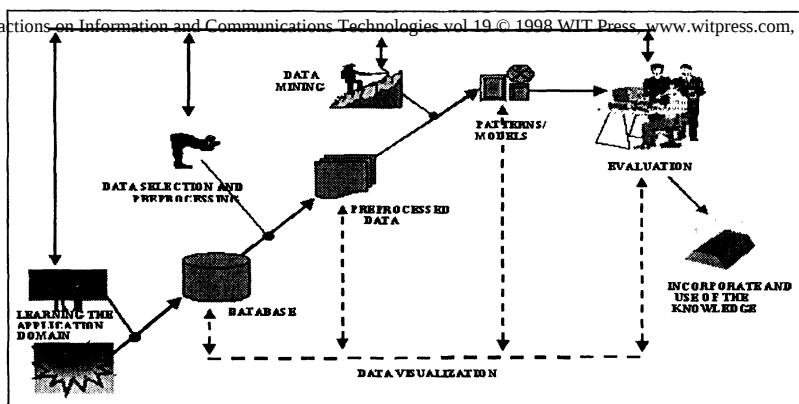


Figure 1. Stages of the Process of Knowledge Discovery in Databases.

The difference between the use of the data visualization and Data Mining methods is that the data visualization makes it possible to directly incorporate perception, human intelligence, calculation and association capacity into the DM process. Some Data Mining techniques (statistical regression, rule based reasoning and artificial neural networks) use mathematical calculations to identify data and relationships. Only data visualization uses human perception as a primary method for the discovery of values. Data visualization and Artificial Intelligence techniques are a powerful combination for Data Mining success.

The visualizations were accomplished with the use of MineSet™, which is a Silicon Graphics Inc. software that incorporates techniques of Data Mining to find interesting situations in databases. This software is made up of control modules, mining tools and powerful visualization tools.

The MineSet™ kernel is a library of Machine Learning algorithms in C++ language, called Machine Learning Library in C++ (MLC++), and is projected to aid in the selection of the best algorithms for certain problems solution Kohavi[8]. The interface between the library and MineSet™ is made through MindUtil™. The visualization tools available in MineSet™ version 2.01 are: Tree Visualizer, Scatter Visualizer, Statistic Visualizer, Map Visualizer and Evidence Visualizer.

Tree Visualizer represents data that has hierarchical relationships and data resulting from Decision Tree Classifiers. With this visualization tool the relationships existent among data can be examined at different hierarchical levels.

Scatter Visualizer can represent data in several dimensions. The data are shown in a grid representing up to three dimensions. A slider



can be used to select specific values along these dimensions, or a path can be traced through these dimensions to make graphical animations. Statistic Visualizer executes a statistic analysis on certain groups of data and generates a window with total, average, maximum and minimum numbers. Map Visualizer represents the data that are geographically distributed in a certain area. This tool also has capacities that allow the user to focus on specific areas and do a more detailed analysis on smaller geographical elements. Evidence Visualizer represents the results obtained from the Evidence Classifier.

Some of the tools described were used to aid in the stages of the Knowledge Discovery in Databases of PMGRN.

4. Techniques of Visualization in the KDD Process

In general, when searching for knowledge, visualization techniques are used in the final stages of the KDD process, aiming at evaluating the knowledge extraction process as a whole. This approach is similar to the software development process which follows the Cascade (or Classical) Model, where testing only starts after all the analysis, project and coding of the software has been done. Changes may be necessary from the initial stages of the process, causing the whole process to become very expensive or even unfeasible Pressman[12]. However, the use of such techniques to aid all the stages of the KDD process can provide a larger control of this process. In the following paragraphs, we show how visualization techniques can cooperate in the accomplishment of the stages of knowledge extraction in databases, using PMGRN as the database.

The PMGRN is based on the Animal Model, which tries to determine the amount of MGT that the parents of bulls and cows manage to transmit to their calves. According to the Experts, MGT and DEP are the most important variables to be considered in the knowledge extraction process. When a calf is born, he immediately is weighed and he/she starts to be part of the Nelore Cattle Registration, where several pieces of data are stored about the program. Later on, the animal is weighed at 120 days, 240 days, 365 days and 550 days after its birth. Other measurements are also made, such as the value of the scrotum perimeter for the bulls and the development of the matrix for the cows. The data of the program is contained in a collection of databases, of which three were considered the most important. They are described below.

Data base 1 refers to the general characteristics of each animal of the program, with a total of 65,000 records. In this database, a large absence of information (missing values), and although there is data from as early as 1958, 85% of the data is concentrated after 1989. Database 2, regarding the males with the best measures of DEPs (when being born, at 120, at 240, at 365 and at 550 days) and MGT, has a total of 2,000 records. Database 3, contains the females with the best measures of DEPs (when being born, at 120, at 240, at 365 and at 550 days) and MGT, with a total of 2,000 records. In bases 2 and 3, more than 98% of the records were complete and there is a history of the best bulls and cows for the years of 1996, 1995 and 1994. There is also a lot of documentation of the bases which is very important for the KDD process.

In the Domain Understanding stage, shown in Figure 1, a total understanding of the scope of the application can be obtained, considering aspects of viability studies, the cost of the application of the process, verification of the amount and of the type of knowledge available before the execution of the process, and conditions and goals of the users, by establishing evaluation criteria for the results of the process and specifying how the extracted knowledge will be applied.

Figures 2 and 3 present the geographical maps that were built to aid the phase of Domain Understanding. In the first graph, represented in Figure 2, data referring to the average weight of the herd was analyzed in the states considered most representative for the program. It can be concluded from this example that the state of São Paulo possesses a larger average than that of the states of Minas Gerais, Goiás and Mato Grosso do Sul. That is to say, the probability that the herds of São Paulo possess better DEP, and consequently MGT, is greater than that of other states.



Figure 2. Visualization for analyzing the regional data in the domain understanding phase.

In Figure 3 a map of some farms belonging to the PMGRN program was built. By using existing animation characteristics of the Map Visualizer tool, the growth of the herds could be analyzed in the main farms. These visualizations, in the form of maps, make it possible to suggest that the place where the cattle is raised can favor the better values of DEP. Thus, place is identified as an important property for the problem.

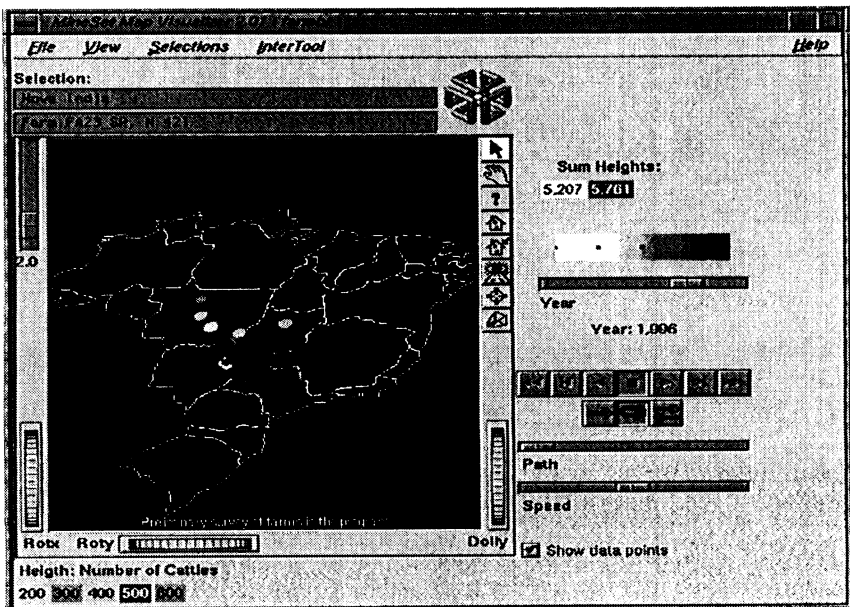


Figure 3. Visualization for analyzing data of farms in the domain understanding phase.

Figure 4 presents the visualizations that were done using the Scatter Visualizer tool. In this specific case, MGT is represented by the size of the entities and by the colors that are represented in the grid, and the coordinates axis represents the Register of the Mother, Register of the Father and Birth Date. In Figure 4(a) all male bulls and their respective MGT are represented. Little can be discovered starting from this visualization window. Indeed, some filter should be applied to the data to see what kind of relationship exists among these variables that can aid in the understanding of the domain. To do this, a tool that is part of Scatter Visualizer was used to filter the data. Figure 4(b) shows the values of MGT of the males that were born between 1984 and 1986 – ten years before the Animal Model was adopted. Note that the entities represented in this visualization are of small and varied sizes, as well as



colors that indicate a low value of MGT. This result contrasts with the result presented in Figure 4(c) where values of MGT are shown for male bulls that were born between 1994 and 1996 - after the adoption of the Animal Model for the Program. The Figure also shows that the entities are homogeneous in terms of size and colors, which indicates that with the adoption of the program the values of MGT improve substantially. This information shows that the viability of knowledge extraction is high due to the positive results of PMGRN.

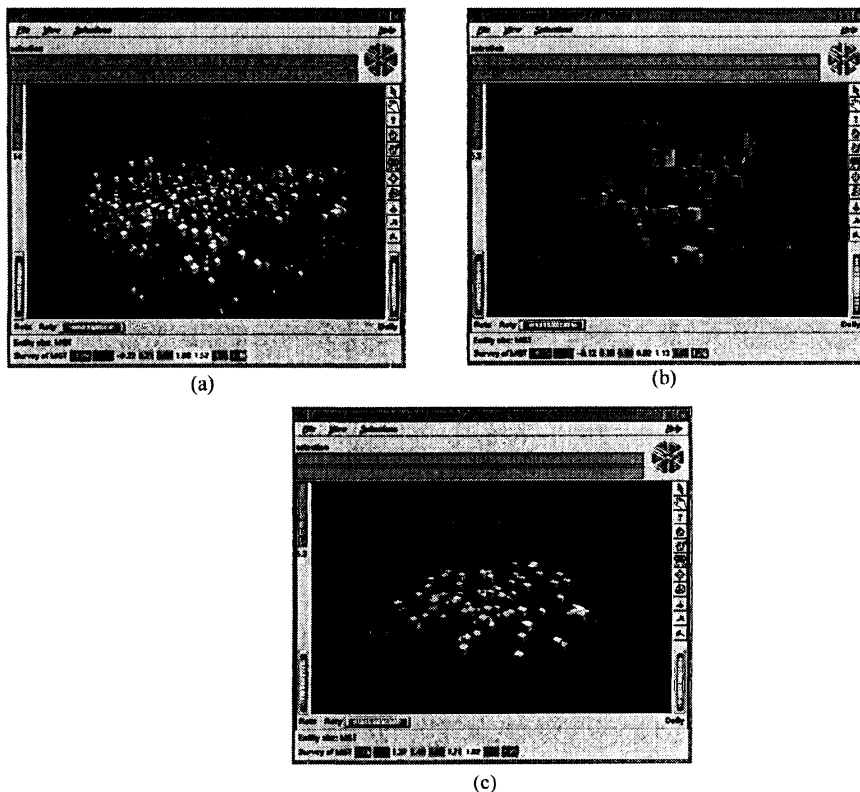


Figure 4. Visualizations of the analysis of MGT for bulls in the domain understanding phase.

The stage of Selection and Preparation of the Data is necessary because direct extraction from large databases can generate an explosive space of search for patterns. Also, most of the algorithms for the extraction of patterns possess limitations with relationship to the volume of data that they can manipulate. This phase concentrates on the selection of a representative sample of the database, and preprocessing and

reduction of the sample, in order to adapt it to the input patterns of Data Mining algorithms.

Different strategies were used for the databases. In Database 1 the data was only considered from 1989 on and several databases were generated: considering all registers that were complete (365 registers); not considering DEPs at 120 days; disregarding those at 365 days and disregarding DEPs at 550 days. In databases 2 and 3 all data was considered. For these two databases, MGT was considered as a major attribute, that is, MGT was considered as an attribute-class.

One way to do the preprocessing of the data is to check for null elements or registers in the system, and then take the appropriate measures to eliminate them. The visualization tools of MineSet™ can treat null registers. In this way it is possible to accomplish visualizations to see how these values can influence the end result of the process. The visualization was accomplished with the same data considered in Figure 4 after having filtered the null data. In Figure 5 the visualization obtained by the tool is shown to analyze the impact of these numbers on the KDD process. A very insignificant number of null data was found in relation to the volume of existent data – as was shown in Figure 4(a).

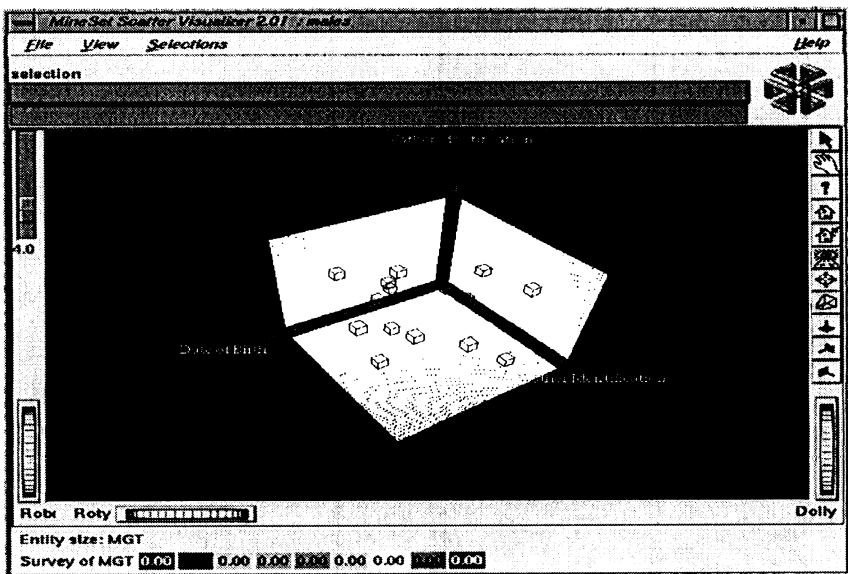


Figure 5. Visualization of the null data of the base of bulls in the phase of selection and preparation of the data.

Data Mining involves the creation of appropriate representation models for the patterns and relationships identified in the data. The result

of these models, after having been evaluated by the Expert and/or Final User, is used to predict the values of attributes defined by the Final User based on new data.

In this stage DEP and MGT were used as the attribute-classes for different analyses. The algorithms CN2 Clark[1] and C4.5 Quilan[13] were used - from the MLC++ library - for the rule extraction process and generation of decision trees. During this stage, some situations were created so as to find patterns in the data. Having done this, all the situations and results were presented to the domain Expert so that he could determine the best ones. From this situation a more detailed work was accomplished with the aid of the available tools.

Figure 6 shows the visualizations that were generated using the Tree Visualizer tool after application of algorithms (inducers) to build classifiers. These classifiers were generated automatically by MineSet™ presenting a training group to it. This training group consists of the PMGRN database register for which the label was determined, that it is the attribute which is to be predicted using the so-called descriptive attributes, which are the attributes used for the prediction. Two visualizations were used to aid in the understanding of how the classifier operates. The first is shown in Figure 6(a) and it only represents the best classifier. This visualization was useful to find the best results obtained in this phase. The second form, shown in Figure 6(b), shows all the classifiers generated during the Data Mining process, seeking to obtain a better understanding of as how the classifier works.

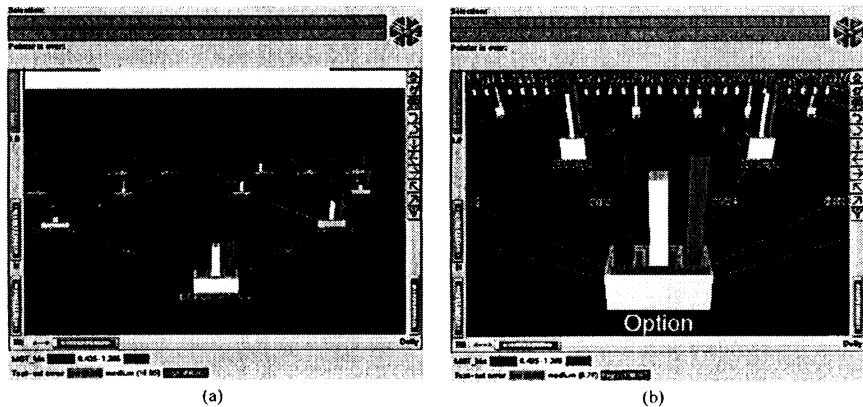


Figure 6. Visualization of the decision trees generated in the Data Mining phase.

The stage of Evaluation of the Knowledge concentrates on the validation of the usefulness of the obtained knowledge judged by the



user. This evaluation of the model is a quite difficult task since it involves, among other aspects, the use of methods (generally statistical) for “filtering” the extracted knowledge, removing redundant patterns and/or the less important ones Fayyad[4]. These methods should be accompanied by visualization techniques to aid the users in the filtering the patterns, as well as in deciding about the usefulness of the extracted knowledge. Also, the criteria for the performance of the process should be observed, considering factors such as the precision and the representation of the extracted knowledge.

The results were presented to the Expert, who noticed how the MGT had been increasing in the last years, a better distinction of the farms that are a part of the program and the importance of the regions in the evolution of MGT. In Figure 7 the participation of MGT in different DEP is shown. In Figure 7(a) the window of the Evidence Visualizer tool is shown that presents all the results obtained for each DEP considered. And in Figure 7(b) the results visualized with the Statistic Visualizer tool.

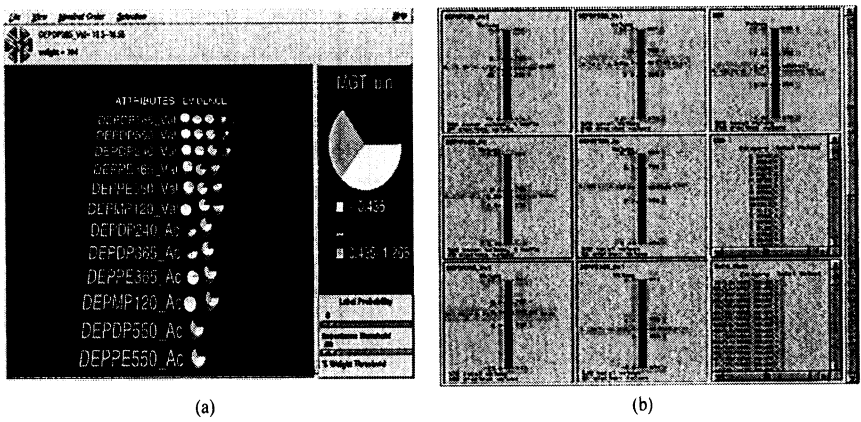


Figure 7. Visualization of the results obtained in the phase of evaluation of the knowledge.

5. Summary

The use of tools capable of representing the data in different visual metaphors is indispensable for getting a good appreciation of the data. Once this has been done, these representations facilitate a better analysis of the data, the information being more explicit in one representation type than in another.



Discovery in Databases. Visualization is useful in the whole process, starting with the learning about the specific domain where visualization can help the communication between Analyst and domain Expert. In the data selection and preprocessing phase, visualization can be used to analyze the preprocessing results. This process improves the refinement of the cleaning of the data.

In the Data Mining step visualization is important to identify the patterns and models in the domain. The final step is the knowledge evaluation. Here the visualization helps the three kinds (categories) of users to consolidate the process because they can easily analyze the objectives and the acquired knowledge. In these latter steps of the knowledge discovery, it is also difficult to evaluate its success.

A very important aspect of the visualizations is that it improves the communication between the Analyst, the Expert and the Final User.

According to the Expert, the prediction of DEP and MGT for cattle and the simulation of mating among the cattle using KDD are some of the first future works that will be accomplished. For this work new attributes should be considered, such as regime (confined, open, mixed), type of feeding (grazing, ration, mixed), area (dry, humid), the characteristics of the father bull (date of birth, weight, MGT, DEP, scrotum perimeter, etc), the characteristics of the mother cow (date of birth, weight, MGT, DEP, matrix and others).

Acknowledgments

This work was partially supported by National Research Council CNPq, FINEP and CAPES and Instituto Mexicano del Petróleo. The authors are also grateful to Dr. Raysildo B. Wolf from Genetics Department of University of São Paulo for kindly proving the database used in this work and to Jaqueline Brigladori Pugliesi and Chandler Wellington Caulkins for helpful corrections to the text.

References

- [1] Clark, P.; Boswell, R. Rule Induction with CN2: Some Recent Improvements, *Proc. 5th European Conference (EWSL 91)*, pp. 151-163, 1991.
- [2] Decker, K. M.; Focardi, S. Technology Overview: A Report on Data Mining. *CSCS-ETH, Swiss Scientific Computer Center*, 1995.



- [3] Fayyad, U. Piatetsky-Shapiro, G.; Smyth, P. The KDD Process for Extracting Useful Knowledge from Volumes of Data. *Communication of the ACM*, vol. 39, nº 11, pp. 27-34, 1996.
- [4] Fayyad, U. Data Mining and Knowledge Discovery: Making Sense Out of Data. *IEEE Expert*, 11 (5), pp. 20-25, 1996.
- [5] Félix, L.C.M.; Rezende, S.O.; Doi, C.Y.; Paula, M.F; Romanato, M.J. MLC++ Biblioteca de Aprendizizado de Máquina em C++. *Relatórios Técnicos ICMC-USP*, No. 72, 1998, 64 pg.
- [6] Inmon, W. H. The Data Warehouse and Data Mining. *Communication of the ACM*, vol. 39, nº 11, pp. 49-50, 1996.
- [7] Kerber R.; Livezey, B.; Simouds, E. A Hybrid System for Data Mining. *Intelligent Hybrid System*, pp. 121-141, 1995.
- [8] Kohavi, R.; Sommerfield D.; Dougherty J. Data Mining using MLC++, a Machine Learning Library in C++, *IEEE*, 1996.
- [9] Li, B. Data Mining Now – *A Survey and Thesis Proposal*. *Department of Computer Science*, New York University, 1996.
- [10] Mannila, H. Data Mining: Machine Learning, Statistic and Databases. *Department of Computer Science*, University of Helsinki, 1997. URL: <http://www.cs.helsinki.fi/~mannila/>
- [11] Oliveira, R.B.T.; Rezende, S.O. Ferramentas de Visualização de Dados do MineSet. *Relatórios Técnicos ICMC-USP*, No. 71, 1998, 47 pg.
- [12] Pressman, R. S. Software Engineering: A Practioner's Approach, *Third Edition, McGraw-Hill Book*, 1994.
- [13] Quilan, J. R. C4.5 Programs for Machine Learning, *Morgan Kaufman Publishers*, 1993.

ⁱ Methods and tools of Data Mining will be called algorithm in this work.